Competing Predictors Segment and Identify Switching Dynamics

J. Kohlmorgen†, K.-R.Müller†, K. Pawelzik‡
† GMD FIRST, Rudower Chaussee 5, 12489 Berlin, Germany
‡ Inst. f. theor. Physik, Universität Frankfurt, 60054 Frankfurt/M., Germany
{Jek;Klaus}@first.gmd.de, Klaus@chaos.uni-frankfurt.dbp.de

1 Introduction

Neural Networks are considered to be a quite general framework for the representation of relations present in data [8] and are frequently used for classification and prediction. An important prerequisite for the successful application of such systems, however, is a certain uniformity of the data. In particular stationarity is often assumed, i.e. that the relations remain constant over time. If, on the contrary, the data set originates from different sources, e.g. because the underlying system **switches** its dynamics, standard approaches like simple multi-layer perceptrons neccessarily must fail. One approach to solve this problem is the inclusion of additional information which helps to disambiguize the data. For instance memory effects can be taken into account by extending the effective input as in time-delayed neural networks [11]. There are, however, problems when applying such systems to data in which unknown sources alternate in time, because a segmentation of the data is required before a **supervised** learning algorithm can be applied. In the present paper we propose a simple **unsupervised** framework which segments a data stream and simultanously identifies the underlying sources. We illustrate this approach with the problem of predicting an unstationary time series in which different dynamics alternate at random. Despite the simplicity of the learning rule, the segmentation of a time series from intermixed chaotic systems or speech data is very precise, thereby leading to accurate classification and nearly optimal prediction.

2 Inert Predictors

Our framework consists of prediction experts which compete for the data. The idea of competing experts has been suggested by Jacobs et. al. [2] to simplify the training task for consistent and unambiguous data sets. Our approach is tailored to the problem of identifying switching dynamics that produces inconsistent data.

Examples for such data are time series from alternating dynamics. The time series can in principle originate from all kinds of systems, like stochastic dynamics or hidden markov models. Phenomena of this kind are e.g. speech, brain data, and systems which switch their attractors [3]. While our framework applies to arbitrary dynamical systems we will here only exemplify it for the case of deterministic chaotic systems and vowel data from speech recognition. We restrict our notation to one-dimensional maps without loss of generality. For higher-dimensional systems states are not necessarily identical with the values of the signal. The state space, however, can be reconstructed [10] from the scalar data and there are methods for the optimal choice of embedding parameters [4].

Let $\{f_l\}, l = 1, \ldots, L$ denote a set of such maps. Then $\{x_t\}, t = 1, \ldots, T$ is a time series from an alternating application $l(t)$ of these maps if $x_{t+1} = f_{l(t)}(x_t)$, i.e. at times t the map $f_{l(t)}$ determines the next state x_{t+1} from the present state x_t.

The goal is now to estimate the parameters of a set of predictors (rsp. experts) $\{\tilde{f}_k\}, k = 1, \ldots, K$ together with a segmentation $\tilde{l}(t)$ only on the basis of the time series, such that the average prediction error is minimal. Thus, to reach an optimal prediction the switching points should be found accurately, i.e. the experts are to identifiy the different sources.

Once an expert $\tilde{f}_k$ performs well with a low error he obtains a competition advantage for nearby data points, i.e. he is supplied with evolutionary inertia to remain in his niche. This can be done by convolving the error $e_k(t) = \mid \tilde{f}_{\tilde{l}(t)}(x_t) - x_{t+1} \mid$ which a predictor k has at time t with a filter kernel. One can furthermore bias the error to represent higher order information, e.g. to represent forbidden and/or probable transitions. In the following examples, however, we only included short term memory and used a moving average $E_k(i) = (\sum_{j=-\eta}^{\eta} e_k(i+j))/(2\eta+1)$. The E_k then are compared in order to determine the winners for times t, which finally are trained on the corresponding parts of the data. The training is started with identically initialized predictors. For the first iteration the training set is divided into subsets of equal size, to ensure a sufficient diverse initialization. In the second and each following training pass a hard competition is carried out: Only the predictor with the smallest $E_k(i)$ is allowed to train on the i-th pattern. This implies, that unnessesary predictors die, i.e they drop out of the learning process.

3 Identifying Switching Dynamics

The performance of our method is illustrated with one-dimensional maps, with the Mackey-Glass equation and with speech data.

In the one-dimensional example an ensemble of six competing predictors is applied to a time series from two maps, the sine map $f_1(x) = \sin(\pi x)$, and the double logistic map $f_2(x) = f_l(f_l(x))$, $f_l(x) = 4x(1-x)$. We use a time series where the dynamics alternates every 50 time steps except for time $t = 300$ (Fig. 1a-d). Furthermore Gaussian noise ($\sigma = 0.1$) is added to the signal to hide its deterministic nature. There is no way to include the overall dynamics of this example into a single map, which would predict an average between the two maps shown in Fig.1(a). As predictors we use radial basis functions that have been suggested by Moody and Darken [6]. Figure 1(b) shows the errors of the six predictors on the time series. The **unsupervised** segmentation in this example identifies the underlying dynamics perfectly, i.e. it is found out **that** and **which** two systems underlie the observed signal, and it is determined **where** the corresponding dynamics are switched.

As a second example we take the formally infinite dimensional Mackey-Glass delay differential equation $dx(t)/dt = -0.1x(t) + \{0.2x(t-\Delta)\}/\{1 + x(t-\Delta)^{10}\}$, where we alternate the delay parameter Δ (see figure caption 1(c) for details). η is successively increased while the training proceeds in order to force the system from some initial coexistence to a stronger competition in the end. Figure 1(d) shows the distribution of winners on the training set after ten iterations. Again the learning method performs well, even for the noisy data set. In both chaotic cases discussed, only the minimal necessary number of predictors survive: each one predicting the corresponding chaotic systems represented by the data.

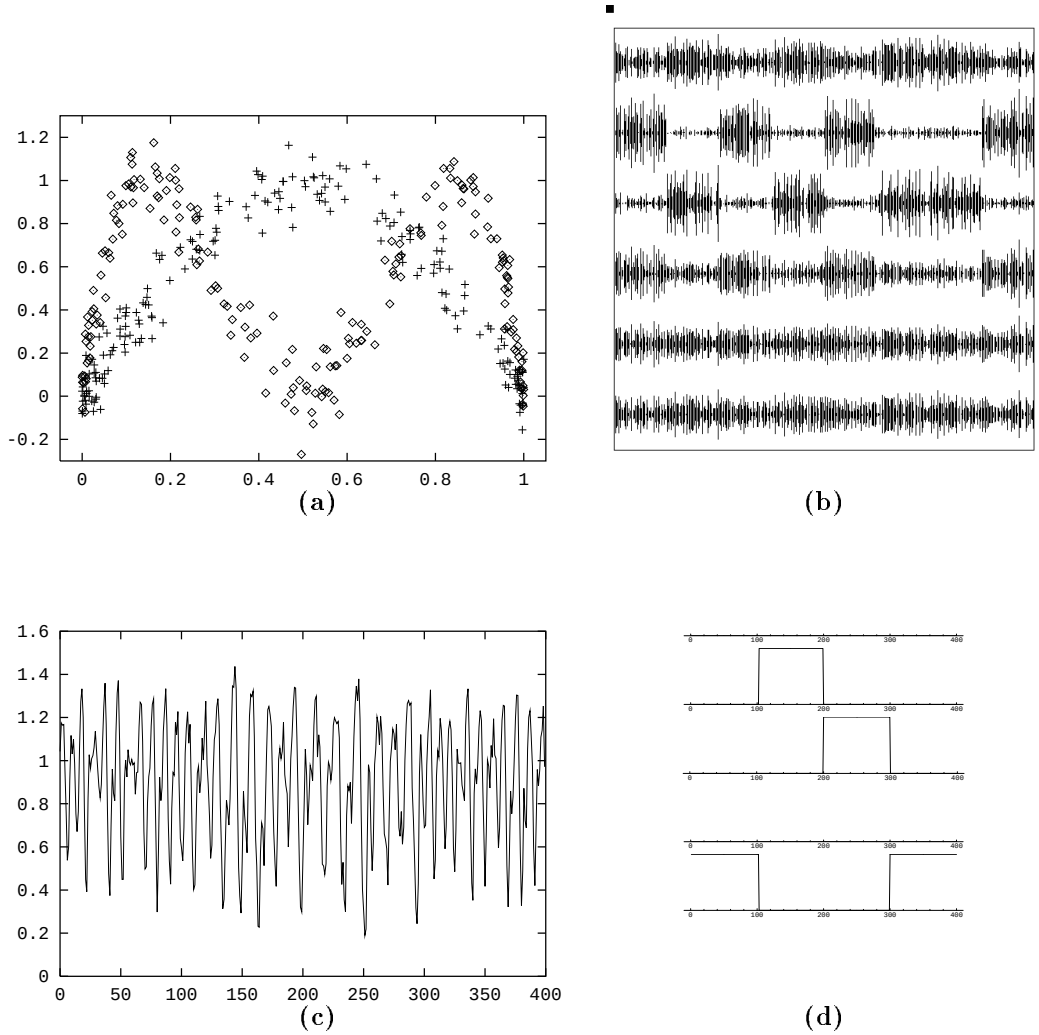

Figure 1: *(a) Return map from a time series of two altenating one-dimensional chaotic maps (sine and double-logistic, with noise of* $\sigma = 0.1$ *added to the time series). Clearly such data cannot be represented by a single function. (b) Errors* e_k *of six predictors after ten iterations of the competition. Choice of the respective winners leads to perfect segmentation (not shown). (c) Non-stationary time series generated by the Mackey-Glass differential equations. The training set consists of 400 samples (sampling rate* $\tau = 6$*). For the first and last 100 samples we choose* $\Delta = 17$*, whereas for the second 100 samples we use* $\Delta = 23$ *and for the third* $\Delta = 30$*. To increase the difficulty of the problem, 5% noise is added at each integration step, thereby turning the system stochastic. Again we use radial basis predictors, however, with an embedding dimension of* $m = 6$ *(cf. [1]). (d) Winner distribution of five predictors on the time series from (c) after ten iterations of the competition.*

The speech data we segmented are 16 dimensional mel-scale FFT coefficients obtained from continuously spoken vowels (AEIOU, single speaker) at a sampling rate of 16kHz. Eight predictors (30 hidden units) are initialized and our algorithm is performed on a 16×4 dimensional input vector since an embedding of $m = 4$ is chosen which captures some memory in the data. Both the learning rate and η are successively increased during training. After training five networks represent A,E,I,O,U respectively, while two nets specialize on silence and one network dies. As a test of the generalization ability we use these nets as predictors for the speech dynamics of a different dataset (same speaker). We observe a clear segmentation on unknown samples, which shows that our method can indeed detect the vowel dynamics in an **unsupervised** manner and generalize properly. Note, that the testing does not involve additional learning. In fact, we achieved a perfect generalization in 14 cases of totally 20 AEIOU samples (that is 70%). All 20 samples are continuously spoken, yet the speed and emphasis is very different. The 6 erroneous generalizations exhibit just a single misidentified subsegment which in most cases reflected the similarity of E with I, and O with U respectively.

4 Summary and Outlook

We presented a framework for the **unsupervised** segmentation of time series. It applies to systems with a nonstationary switching dynamics, a phenomenon which is observed in many natural signals as e.g. in speech and in brain data [7]. The method is based on hard competition together with a tendency to take advantage of neighborhoods in time (evolutionary inertia). We applied this general idea to time series from alternating maps, switching differential equations and vowel data. It was demonstrated that our approach leads to nearly perfect segmentation even in the presence of noise. Note, that the goal at this point of our study is not to convince the reader to use our unsupervised method in his speech recognition system, but to demonstrate its remarkable universality and simplicity on a number of different applications with switching dynamics.

Acknowledgement: Prof. Waibel's group supplied the preprocessed speech data. Valuable discussions there are also gratefully acknowledged.

References

[1] Casdagli, M., Physica D **35**, 335-356 (1989).
[2] Jacobs, R.A. et.al., Neur. Comp. **3**, 79-87 (1991).
[3] Kaneko, K., Phys. Rev. Lett. **63**, 219 (1989).
[4] Liebert, W., Pawelzik, K., Schuster, H.G., Europhys. Lett. **14**, 521 (1991).
[5] Mackey, M., Glass, L., Science **197**, 287 (1977).
[6] Moody, J., Darken, C., Neural Computation **1**, 281-294 (1989).
[7] Pawelzik, K. et.al., Proc. of NIPS 92, Morgan Kauffmann, 977-984 (1993).
[8] Rumelhart, D.E., McClelland, J.L., MIT Press, Cam. Mass. (1984).
[9] Schuster, H.G., *Deterministic Chaos*, 2nd Ed., Physik Verl., Weinheim, (1988).
[10] Takens, F., Springer Lecture Notes in Mathematics, **898**, 366 (1981).
[11] Waibel, A. et.al., IEEE Trans. on ASSP, **37**, 328 (1989).